PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Long-term prediction of major coronary or ischemic stroke event in a low-incidence Southern European population: model development and evaluation of clinical utility
AUTHORS	Veronesi, Giovanni; Gianfagna, Francesco; Chambless, Lloyd; Giampaoli, Simona; Mancia, Guiseppe; Cesana, Giancarlo; Ferrario, Marco

VERSION 1 - REVIEW

REVIEWER	Collins, Gary University of Oxford, Centre for Statistics in Medicine
REVIEW RETURNED	23-Aug-2013

THE STUDY	No information on inclusion/exclusion criteria. The authors are pooling data from 4 studies and it's unclear whether all data from all studies is being used.
GENERAL COMMENTS	The authors describe the development of a new model for predicting the 20-year risk of developing CVD. The manuscript is generally well written and the methods entirely appropriate for a study developing a multivariable prediction model. My comments are therefore relatively minor. Title: The model is developed using data from Italy; I would therefore suggest the authors change the 'European population' text to make it clearer the model is intended to be used in Italy. Whilst it would be useful if the model worked in other European countries, this is not guaranteed and would require external validation on data form other countries to confirm this. The authors are pooling data from 4 independent population surveys. Some reassurance that pooling these datasets together is sensible would be useful. A table (supplementary) of characteristics (like the current table 1) for the 4 cohorts would be useful. Page 7: Is the use of anti-hypertensive treatment in the last 2 weeks a sensible approach? Isn't use in the previous 6 months more realistic? I would also like to have more information on inclusion/exclusion criteria. E.g ranges of any continuous predictors. Table 1 states patients were aged between 35 and 69, ok. But what about the systolic blood pressure, HDL and total cholesterol. Applying a model on someone outside these ranges is extrapolation. So indicating exactly who the model is intended for is important to know. Also I have no idea if all patients from all the 4 surveys is being used. Some clarification in the methods that the authors fit a full model and did not do any pre-selection of predictors (or that's how it reads, but

	needs confirming). Unless the authors have done something else, which will need describing. Page 8: I do fail to see the rationale of categorising total and HDL cholesterol. Particularly when age and systolic blood pressure are retained as continuous. Just because other studies decided they would be happy in losing information by categorising continuous predictor doesn't mean the authors should persist with it. Page 8: the authors make reference to quintiles. This is technically incorrect and should be labelled 'fifths', can the authors replace this throughout. Quintiles, of which there are 4 or 6 depending on how one sees the lower and upper bounds, and are merely cut points to create five equally sized groups. See Altman DG, Bland JM. Statistics Notes: Quartiles, quintiles, centiles, and other quantiles. BMJ 1994; 309:996. Page 8: The authors evaluate calibration using the Gronnesby-Borgan GOF. I would also like to see a calibration plot of the prediction models. A p-value give me no indication on any miscalibration or magnitude. Page 8: I don't fully understand 'We then compared the AUC of both models, by considering bootstrapped confidence intervals for the difference in the betas' Can the authors clarify, how they are doing the bootstrapping. Did (if my assumption on how the built the model is correct) the authors for the bootstrapping take bootstrap resamples, fit a full model, evaluate it on both the original data and bootstrap sample, take the difference, repeat this a number of times and then get the mean of this different to get the optimism? I would also like to know the number of 10 year events.

REVIEWER	Tzoulaki, Ioanna Imperial College London, Epidemiology and Public Health
REVIEW RETURNED	19-Sep-2013

GENERAL COMMENTS	This article presents the developemnt of a risk prediction model for CVD over 20 years period. I ofund teh methods used clear and appropraite and well-reported. I am however thoughtful of the nessecity to develop thsi model. We already have a 10 year model and teh FRS recently publihsed a 30 year one. Why do we need anothr model? It is well described that we produce more models than we need and we do not validate our models. In adittion, the authors choose to compare the 20 year model to the 10 year CUORE one. Why did teh authors use teh CUORE over teh more widely used FRS and SCORE? It woudl be good to add them as well and comapre their predictive performance. Also it is unlcear how the CUOTE 10 year was calculated. Did the authors cenceor all analyses at 10 years? and did they use the published coefficients or CUORE? The net benefit analysis is interesting but the weighst used shoudl be mentioned and explained. Also, the authors should consider the long term side effects of statins (in adittion to costs) is statins are presecribed based on 20 year risk and therefore for longer periods of time. I was expectingto see Hosmer Lemeshow statistic sin your

	calibration analysis, is there any reason that has not been used? Throughout your manuscrip and in the conclusions you shoudl emphasise that a major limitation of thsi study is teh fact that teh model has not been externally validated and only cross-validation has taken place. The value of the model is still unknown.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Dr Gary Collins, NDORMS Senior Research Fellow, University of Oxford

No information on inclusion/exclusion criteria. The authors are pooling data from 4 studies and it's unclear whether all data from all studies is being used.

R. These arguments were further developed by the reviewer below.

The authors describe the development of a new model for predicting the 20-year risk of developing CVD. The manuscript is generally well written and the methods entirely appropriate for a study developing a multivariable prediction model. My comments are therefore relatively minor.

Title: The model is developed using data from Italy; I would therefore suggest the authors change the 'European population' text to make it clearer the model is intended to be used in Italy. Whilst it would be useful if the model worked in other European countries, this is not guaranteed and would require external validation on data form other countries to confirm this.

R. We respectfully disagree with the reviewer. The current ESC guidelines and the SCORE Project risk charts differentiate low- from high-risk Countries; the risk gradient follows a well-known geographical pattern. Our contribution aims to make the scientific community aware of the clinical utility of a long-term risk prediction model also in a low-incidence population. For these reasons, we added the term "Southern" to the title. Recommendations and limitations on the use of this model in a different Country have been strengthened in the discussion section.

The authors are pooling data from 4 independent population surveys. Some reassurance that pooling these datasets together is sensible would be useful. A table (supplementary) of characteristics (like the current table 1) for the 4 cohorts would be useful.

R. As part of model development analysis, a "cohort" term was added to the full model, to test for any unwanted study effect (3 df test). The term was not statistically significant, in men (p-value: 0.21) and in women (p-value: 0.45). A sentence was added at the beginning of the statistical analysis, page 6. In addition, the table suggested by the reviewer has been added as Table S2 in the supplementary material.

Page 7: Is the use of anti-hypertensive treatment in the last 2 weeks a sensible approach? Isn't use in the previous 6 months more realistic?

R. In population-based studies this is a standard, from the WHO-MONICA Project experience on. The same information was used in the CUORE Project (see Ferrario et al, Int J Epidemiol 2005, reference 13 in our paper).

I would also like to have more information on inclusion/exclusion criteria. E.g ranges of any continuous predictors. Table 1 states patients were aged between 35 and 69, ok. But what about the systolic blood pressure, HDL and total cholesterol. Applying a model on someone outside these ranges is extrapolation. So indicating exactly who the model is intended for is important to know. Also

I have no idea if all patients from all the 4 surveys is being used.

R. The ranges for continuous risk factors have added in the footnote of Table S1. As mentioned, we considered for this analysis only subjects in the 35-69 age range without previous CVD event at baseline (paragraph 1 in the results section).

Some clarification in the methods that the authors fit a full model and did not do any pre-selection of predictors (or that's how it reads, but needs confirming). Unless the authors have done something else, which will need describing.

R. The risk factors included in the long-term model were part of the CUORE Project 10-year risk equation for the Italian population (Ferrario et al, Int J Epidemiol 2005, reference 13), which was developed by the same research team. Furthermore, the same factors have been standard in CVD risk scores around the world for many years (Chambless LE et al, J Clin Epidemiol 2003 [ref 4]; D'Agostino RB et al, Circulation 2008 [ref 3]). In a different paper we will address the additional contribution of other risk factors discussed in European and US guidelines for primary prevention, including selected biomarkers, family history of coronary heart disease and social status. This was better clarified in the statistical analysis section.

Page 8: I do fail to see the rationale of categorising total and HDL cholesterol. Particularly when age and systolic blood pressure are retained as continuous. Just because other studies decided they would be happy in losing information by categorising continuous predictor doesn't mean the authors should persist with it.

R. As part of the preliminary analyses we looked at event rates by standard cholesterol categories. The risk did not seem to increase linearly, especially among women; this can be noted also from the beta coefficients reported in Table S1 (supplementary material). Thus, there is a trade-off between including the more complex continuous factors to capture the non-linearity, or staying with the categorical approach that is nearly standard in CVD risk prediction. In addition, categorizing both total- and HDL-cholesterol led to an improvement in model calibration, in both genders, over the use of continuous linear terms. Therefore, although we agree with the reviewer that categorizing a continuous variable is not always a good choice, as also is not the simply use of linear continuous factors, we felt that categorization was the best option in our particular case. The use of standard categories for cholesterol, already adopted in studies similar to ours, should limit the risk of overfitting. This was briefly documented at the beginning of the "statistical analysis" section (page 6).

Page 8: the authors make reference to quintiles. This is technically incorrect and should be labelled 'fifths', can the authors replace this throughout. Quintiles, of which there are 4 or 6 depending on how one sees the lower and upper bounds, and are merely cut points to create five equally sized groups. See
Altman DG, Bland JM. Statistics Notes: Quartiles, quintiles, centiles, and other quantiles. BMJ 1994; 309:996.

R. We changed the word "quintile" into "fifth" (page 8).

Page 8: The authors evaluate calibration using the Gronnesby-Borgan GOF. I would also like to see a calibration plot of the prediction models. A p-value give me no indication on any miscalibration or magnitude.

R. The original submission did not include the calibration plot due to space concerns. We added the figure as supplementary material (Figure S1). The reference to the plot was added in the text (see page 8).

Page 8: I don't fully understand 'We then compared the AUC of both models, by considering bootstrapped confidence intervals for the difference in the betas' Can the authors clarify, how they are doing the bootstrapping. Did (if my assumption on how the built the model is correct) the authors for the bootstrapping take bootstrap resamples, fit a full model, evaluate it on both the original data and bootstrap sample, take the difference, repeat this a number of times and then get the mean of this different to get the optimism?

R. The last part of the sentence was added by mistake. The correct sentence should have been: "We then compared the AUC of the two models by looking at their respective bootstrapped confidence intervals." We modified the text accordingly.
We confirm that the procedure described by the reviewer in his comment is what we did to estimate the optimism in discrimination.

I would also like to know the number of 10 year events.

R. The number of events for the 10-year analysis is: 234 for men and 79 for women. A sentence was added in the statistical analysis section, page 6.

Reviewer: Ioanna Tzoulaki, Lecturer in Epidemiology, Imperial college, UK

This article presents the developemnt of a risk prediction model for CVD over 20 years period. I ofund teh methods used clear and appropraite and well-reported. I am however thoughtful of the nessecity to develop thsi model. We already have a 10 year model and teh FRS recently publihsed a 30 year one. Why do we need anothr model? It is well described that we produce more models than we need and we do not validate our models.

R. Recent European and US guidelines for CVD primary prevention are moving forward the concept of long-term risk, as a complement for the shor-term models. Our research team previously showed that a prediction model developed on a high-incidence population like the US one does not work well in a low-incidence population even after recalibration (Ferrario et al, Int J Epidemiol 2005, reference 13). This is clearly stated in the introduction (page 3, "To this extent,…high-incidence countries").
In addition, threshold values for risk stratification should be re-evaluated before being applied in a different population (Collins GS et al, BMJ 2012, reference 16). Our paper underlines the importance of a clinical utility analysis as part of model development and validation, along with all the other statistical considerations.

In adittion, the authors choose to compare the 20 year model to the 10 year CUORE one. Why did teh authors use teh CUORE over teh more widely used FRS and SCORE? It woudl be good to add them as well and comapre their predictive performance.
Also it is unlcear how the CUOTE 10 year was calculated. Did the authors cenceor all analyses at 10 years? and did they use the published coefficients or CUORE?

R. As part of model development, we verified whether a single measurement of risk factors could still be predictive of events at 20-year distance from measurement, with no significant drop in discrimination with respect to short-term prediction. We did not compare our long-term model with any specific published equation, including the CUORE one. We already mentioned above the concerns about using the FRS in Italy; while we could not have compared our model with the SCORE one since non-fatal events were part of the endpoint in our analysis.
At the light of this comment, we modified the beginning of the statistical analysis section at page 6 to avoid any possible misinterpretation.

The net benefit analysis is interesting but the weighst used shoudl be mentioned and explained. Also, the authors should consider the long term side effects of statins (in adittion to costs) is statins are presecribed based on 20 year risk and therefore for longer periods of time.

R. The Net Benefit “per se” cannot be referred to any single treatment nor takes the cost-effectiveness of any treatment into account. The weights are fixed at any given value of predicted risk (for example, at 20% risk threshold, w=0.25 and so on), reflecting the odd of the selected threshold for high risk designation. Data on cost-effectiveness of long-term statin use are scanty, if any; our reference [30] is mainly referring to a 10-year period. We specified this aspect more clearly to the sentence at page 11 (“Despite the lowering costs of statin treatment…over a 10-year period [30]”).

I was expecting to see Hosmer Lemeshow statistic sin your calibration analysis, is there any reason that has not been used?

R. The Hosmer-Lemeshow statistic fails in adequately considering censorship, as it is based on observed counts. The Grønnesby-Bogan goodness-of-fit statistic instead takes censorship into account, as it considers Kaplan-Meier estimates of observed risk. For more details see May S. et al, Lifetime Data Anal 1998 [reference 23 in our paper].

Throughout your manuscrip and in the conclusions you shoudl emphasise that a major limitation of thsi study is teh fact that teh model has not been externally validated and only cross-validation has taken place. The value of the model is still unknown.

R. The external validation for long-term models is in general an issue, since it requires cohort studies with 20+ years of follow-up and a consistent event definition and validation over time. We can only provide a rigorous cross-validation, as in the Framingham 30-year risk score (see Pencina et al, Circulation 2009, reference 10 in our paper). We have planned external validation using other population-based cohorts from different part of Italy. We state this limitation more clearly at page 12 (“A major limitation…to different contexts”) as well as at the corresponding sentence in the “Article summary” section (page 14).